Meta-omics approach to explore microbial community of the wood ant *Formica exsecta*

Kishor Dhaygude

Organismal and Evolutionary Biology Research Programme,
Faculty of Biological and environmental sciences, University of Helsinki.

Academic dissertation

To be presented, with permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, for public examination in Porthania-PIII in Yliopistonkatu 3, Helsinki on December 14th, 2019 at 12.15

Helsinki 2019

Supervised by:

Prof. Liselotte Sundström
Organismal and Evolutionary Biology
University of Helsinki, Finland

Dr. Juan Galarza
Department of Biological and Environmental Sciences
University of Jyväskylä, Finland

Dr. Helena Johansson
Organismal and Evolutionary Biology
University of Helsinki, Finland

Reviewed by:

Prof. Jürgen Gadau
Institute for Evolution and Biodiversity
University of Münster, Germany

Dr. Romain Libbrecht
Institute of Organismic and Molecular Evolution
University Mainz, Germany

Examined by:

Dr. Lumi Viljakainen
Department of Ecology and Genetics
University of Oulu, Finland

Custos:

Prof. Liselotte Sundström
Organismal and Evolutionary Biology
University of Helsinki, Finland

Thesis advisory committee:

Prof. Pekka Pamilo
Organismal and Evolutionary Biology
University of Helsinki, Finland

Dr. Ari Löytynoja
Institute of Biotechnology
University of Helsinki, Finland

Cover illustration: Kishor Dhaygude 2019

ISBN 978-951-51-5599-3 (paperback)
ISBN 978-951-51-5600-6 (PDF)
http://ethesis.helsinki.fi
Unigrafia Helsinki 2019

The Faculty of Biological and Environmental Sciences uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

"If you want to shine like a sun, first burn like a sun".

- A. P. J. Abdul Kalam

Contents

Abstract 6

Introduction 7
- Exploration of natural ecosystems and microbiomes 7
- Next Generation Sequencing application in non-model organisms 7
- *Formica exsecta* and its interactions with microbes 10

Aims of the Study 11

Material and Methods 12
- Study Species 12
  - *Formica exsecta* and other *Formica* species 12
- Meta-Transcriptomics Analysis 13
  - RNA extractions & sequencing 13
  - Transcriptome assembly 13
- Microbiome Analysis 14
  - Identification of microbiota 14
- Meta-Genomics Analysis 14
  - DNA extraction & sequencing 14
  - Genome assembly 15
  - Genome Annotation 15
- Phylogenetics Analysis 16
  - Bacteria Phylogeny 16
  - Virus phylogeny 16

Main results and their interpretation 17
- Insights from transcriptome profiling of *F. exsecta* 17
  - Caste-specific expressed transcripts 19
  - Expression profile of gene families 19
  - Microbial community diversity 20
- Insights from *F. exsecta* genome 20
  - Horizontal gene transfers, and functional novelty 22
  - Valuable resources for future comparative genomics 22
- Insights from *w*Fex genome 23
- Nature of FeV1, FeV2, and FeV4 viruses 24

Conclusions 26

Acknowledgements 27

References 29

Appendices
- Chapters I–IV

This thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I. **Dhaygude K**, Trontti K, Paviala J, Morandin C, Wheat C, Sundström L, Helanterä H. 2017. Transcriptome sequencing reveals high isoform diversity in the ant *Formica exsecta*. PeerJ 5:e3998. DOI: 10.7717/peerj.3998.

II. **Dhaygude K**, Nair A, Johansson H, Wurm Y, Sundström L. 2019. The first draft genomes of the ant *Formica exsecta*, and its *Wolbachia* endosymbiont reveal extensive gene transfer from endosymbiont to host. BMC Genomics. DOI: 10.1186/s12864-019-5665-6.

III. Johansson H*, **Dhaygude K*,** Lindström, S, Helanterä H, Sundström L, and Trontti K. 2013. A Metatranscriptomic approach to the identification of microbiota associated with the ant *Formica exsecta*. PLoS One 8, e79777. doi:10.1371/journal.pone.0079777

IV. **Dhaygude K**, Johansson H, Kulmuni J, Sundström L. 2019. Genome organization and molecular characterization of the three *Formica exsecta* viruses—FeV1, FeV2 and FeV4. PeerJ 6:e6216. DOI: 10.7717/peerj.6216.

***** These authors contributed equally to this work

Authors Contributions

	I	**II**	**III**	**IV**
Original idea	KD, HH	KD, LS	KD, HJ	KD
Study design	KD, HH	KD	KD, HJ	KD
Field work	KD, HH, CM, LS	KD, LS	KD, SL, LS	KD, LS, HJ
Methods	KD, KT, JP	KD, AN	KD, HJ, KT	KD, HJ, JK
Data analysis	KD	KD, AN	KD, HJ	KD
Manuscript	KD, HH, KT, LS	KD, AN, HJ, YW, LS	KD, HJ, LS, HH	KD, HJ, JK, LS

Kishor Dhaygude (KD), Abhilash Nair (AN), Claire Morandin (CM), Heikki Helanterä (HH), Helena Johansson (HJ), Jenni Paviala (JP), Jonna Kulmuni (JK), Kalle Trontti (KT), Liselotte Sundström (LS)., Stafva Lindström (SL), Yannick Wurm (YW)

Abstract

The majority of the planet's biological diversity comprises of diverse microorganisms, including large communities of insects. It is only through symbiotic, pathogenic and vectoring association, a diverse relationship between the microorganisms and the insects can be established. In spite of having an independent interaction, microorganisms are expected to fulfill the important roles of insect nutrition, reproduction, development, as well as behavioral resistance to pathogen colonization. So to understand the molecular diversity, population structure, and ecological importance of the majority of microorganisms, it is very essential to discover and characterize these microbial communities. The multi-omics approaches have the potential of in-depth screening of microorganisms as well as answering some fundamental microbial ecology questions. So, multi- omics approaches and bioinformatic analysis are considered as the powerful tool to study the non-model microbes and ultimately to study the composition and function of dynamic microbial communities. In spite of these, the microbial community largely remains unknown to the domain of social insects.

This thesis majorly utilizes the multi-omics approaches for demonstrating the dynamic interplay between host and microbes. On the basis of the observational study it has been found that pathogenic and natural microbial community are associated with ant *Formica exsecta*. The findings included members of several endogenous bacterial phyla, such as *Wolbachia*, two obligate endogenous and possibly entomopathogenic fungi, as well as complete genomes of three novel RNA viruses belonging to the classes of Iflaviridae, Dicistroviridae and Mononegavirales. In this thesis, RNA sequencing data for the ant *F. exsecta* constructed from the samples of several life stages of both sexes as well as female castes of queens and workers to maximize the representation of expressed genes. Additionally, for the first time the horizontal gene transfer is demonstrated in this thesis from *Wolbachia* endosymbiont to host *F. exsecta* ant genome and at the same time the process of releasing of the first genome of *Wolbachia* endosymbiont from ant species. Moreover, the focus of thesis is on genome organization and molecular characterization of the three *F. exsecta* viruses and at the same time explaining the viral transmission in other related ant species.

By adopting the advantages of the power of genomic technologies, this thesis tries to provide new insights into the host and microbe interactions, and the evolution of host-parasite genomes in a more general framework. However, in general the studies of this thesis provide useful information, guidelines and resources for social insects and genomics research.

Introduction

Exploration of natural ecosystems and microbiomes

Insects and soil environmental niches are believed to be the preeminent sources of novel biomolecules (Krishnan et al., 2014). Based on the sheer number of species, presence in in ecological habitats, and biomass, insects are undoubtedly the most diverse and abundant animal group (Engel & Moran, 2013; Krishnan et al., 2014). The ecological success of insects is linked to their key relationships with favorable symbiotic microorganisms, influencing their physiology, ecology, and evolution. In several ways, symbiotic microbiomes are favorable to their insect hosts, including dietary supplementation, tolerance to environmental perturbations, and maintenance and/or enhancement of host immune system homeostasis (Weiss & Aksoy, 2011). Insects often depend on nutritionally restricted diets, such as sterile blood, plant juices, or woody material, which requires the presence of mutualistic endosymbionts for the provisioning of essential nutrients by aiding the absorption of these food materials (Douglas, 2011). In addition to obligate endosymbionts, many insects harbor bacteria that are not essential for their survival or fecundity. Such symbionts can influence reproductive traits in the insect hosts, including male-killing, feminization, parthenogenesis, or cytoplasmic incompatibility (Hughes & Rasgon, 2012; Correa & Ballard, 2016). For example, *Wolbachia* are endosymbiotic bacteria of arthropods and nematodes, that may influence the reproduction of infected hosts by causing sterility. Further, many studies have shown that gut-associated microbial communities can influence the immune functions in their insect hosts via alimentary metabolism (Krishnan et al., 2014; Prasad et al., 2018). These gut microbes are capable of producing various cellulolytic and ligininolytic enzymes that can break down the most abundant biological macromolecules and facilitate their digestion (Prasad et al., 2018).

Next Generation Sequencing in non-model organisms

The diversity of microbial communities has been well-studied for model organisms, yet advancement in sequencing approaches provides new tools to supplement culture-based approaches for identifying microbiota. Earlier culture-based methods were used for screening bacterial community based on low throughput sequencing of the bacterial 16S ribosomal rRNA gene (Schmidt, DeLong & Pace, 1991; Tringe & Hugenholtz, 2008).

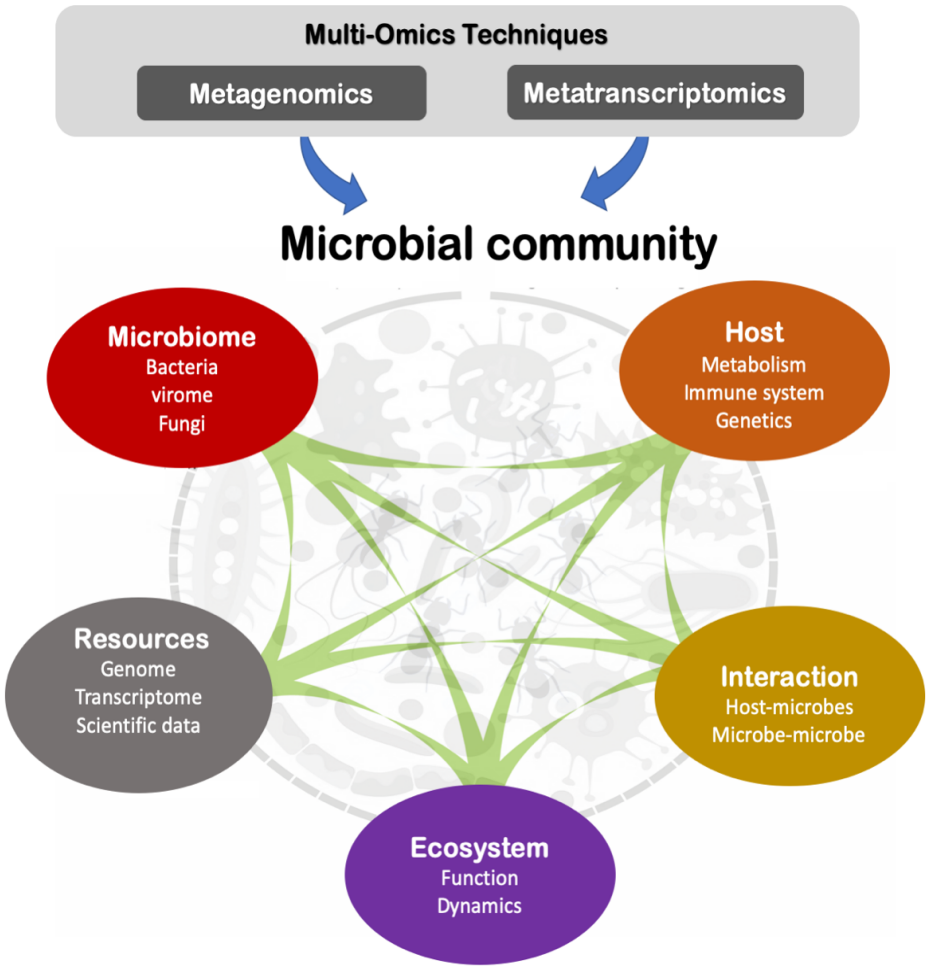

Figure 1: Graphical picture represents interaction in microbial community and emerging techniques for their exploration. Microbial communities are complex biological entities interacting with the environment, host organisms, and transient microbes.

More recently, genome-wide sequencing approaches, such as metagenomics and metatranscriptomics (Figure 1), have further expanded the range of available experimental tools for studying the microbiome. Microbiome is a term that describes the collective genome of the microorganisms in a community, including bacteria, archaea, viruses, and fungi. According to terminology metagenomics is a branch to genomics, in which studies are conducted on collective genome/genes sequencing of

genetic material extracted from environmental samples. Metatranscriptomics (cDNA sequencing) is used to assess gene expression in individuals or their microbial communities. These multi-omics approaches allow the identification of genes, transcripts, and eventually proteins from thousands of microbes, and their hosts, paving the way for analyzing the biological function and host-microbial interactions (Figure 1). Hence high-throughput sequencing technologies are revolutionizing the life sciences especially when we can use this technology to shift the focus from model organisms to non-model organisms. Fortunately, the continual decrease in the cost of sequencing has now made it feasible to determine a genome or transcriptome sequence also for non-model organisms. During the past few years, public databases such as NCBI, EMBL, DDBJ, and others, provide fundamental data significant to biological research, as many new genome sequences from non-model organisms have become available (Russell et al., 2017). We know very little about non-model microbes, even though they constitute 99% of the Earth's biosphere (Wilson, 1971). The multi-omics approach has become a high priority in the field of microbial ecology; with Next Generation Sequencing (NGS) it is now possible to generate and process massive amounts of data for diverse omics studies such as meta-genomics, meta-transcriptomic etc. These approaches have been used to answer broader questions in ecology, including exploring the unknown microbial diversity harboured by host organisms (Sogin et al., 2006; Ge et al., 2012; Kim, Whon & Bae, 2013; Gibbons & Gilbert, 2015; Staley & Sadowsky, 2016; Fanning et al., 2017). So, with meta-genomic approaches one can resolve both the genome characteristics of free living non-model organisms, and identify and classify the microbiota for a specific species group. This represents a new avenue of study in the fundamental research domain of ecology and evolution.

Social Hymenoptera (ants, bees and wasps) rank among the ecologically most important taxa, due to their ecological dominance, and key stone status in many terrestrial habitats (Conte & Hefetz, 2008). Representatives of social Hymenoptera play a key role in pollinating both wild and crop plants. A cornerstone in their ecological success is the emergence of specialized castes, workers that specialize on brood care and foraging, and sexual males and females that reproduce. Nonetheless, social Hymenoptera, like all other organisms, are exposed to, and challenged by parasites, ranging from predatory mites and parasitic flies to fungi, viruses, bacteria, and protozoa (Schmid-Hempel, 1995). Despite its benefits and successes, social life increases vulnerability to pathogens, as the close contact between individuals in populous nests, allows the rapid spread of disease. However to overcome these types of challenges, the social insects fight back with their personal immune system as well as social immunity (Cremer, Armitage & Schmid-Hempel, 2007; Meunier, 2015).

Formica exsecta and its interactions with microbes

Due to their social structure with closely related individuals (Hölldobler & Wilson, 1990), ants offer unique opportunities for using multi-omics approaches in exploring the ecology and evolution of advanced societies (Deslippe, 2010). Surprisingly, only 20 completely sequenced ant genomes have been published during the last seven years (Boomsma et al., 2017), and among these some key sub-families are absent. One such omission is the sub-family Formicinae, and the genus *Formica*, a genus that comprise over 160 species, including the well-known mound-building wood ant species, which have been extensively studied in the context of social evolution (Sundström, Chapuisat & Keller, 1996; Brown, Liautard & Keller, 2003; Sundström, Keller & Chapuisat, 2003a; Vitikainen, Haag-Liautard & Sundström, 2011; Haapaniemi & Pamilo, 2012).

The genus *Formica* is a dominant species in boreal forest ecosystems (Hölldobler & Wilson, 1990). The acquisition of genomic and transcriptomic resources for the genus *Formica* allows a better understanding of the genetic basis of social traits. Comparative genome analysis can aid identification of lineage-specific genes, which may generate evolutionary novelty at the phenotypic level. Also, it will help explain the genomic underpinnings of key innovations that globally characterize the ants as a group, as well as those that characterize the distinct subgroups. Genome sequencing also offers a route toward quantifying the overall role of adaptive evolution, and how this varies among lineages. In addition to genomic data, transcriptome data are important for interpreting the functional elements of the genome, and revealing the molecular constituents of the cells and tissues, including their role in the development of various traits. Transcriptomics aims to catalogue transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, for splicing patterns and quantifying the changing expression levels of each transcript during development and under different conditions.

Aims of the Study

The aim of thesis is to sequence and assemble the genome of the ant *Formica exsecta*, to investigate the microbial communities, and provide molecular insights into an intimate host-parasite interaction by utilizing the genomics resources and associated functional data. I give a brief overview of the four Chapters included in this thesis below:

- The aim of Chapter I is to characterize the transcriptome of *F. exsecta* (including data from multiple castes, life stages differing in many physiological processes and sexes) to provide a good representation of expressed genes with their possible isoforms, which are potentially useful as a valuable genomic resource for future studies on social insects. An additional aim is to identify the caste specific expression profile of *F. exsecta*.

- The aim of Chapter II is to examine the relationship between *F. exsecta* and its *Wolbachia* endosymbiont. More specifically, the aims are to produce the first genome for the ant genus *Formica* and its *Wolbachia* endosymbiont. I identify the horizontally transferred genetic elements from *Wolbachia* in the genome of the ant *F. exsecta*, along with describing the genomic organization of any such elements.

- In Chapter III, the aim is to look into the potential intracellular pathogens and symbionts including RNA/DNA viruses, bacteria, and fungi; as well as other microbiota closely associated with the ants. This chapter revealed the biological interactions between the ant *F. exsecta,* and the micro- and macrobiotic communities that it potentially encounters.

- In Chapter IV, the focus is on RNA viruses infecting the ant *F. exsecta* to understand their primary genome structure and organization, and taxonomy. Additionally, the aim is to examine the molecular evolution of viruses in *Formica* species.

Material and Methods

This thesis contributes to our understanding of functional genomics of host and its microbial association. Therefore, it is necessary to combine multi-omics methods with traditional molecular and bioinformatics analysis, such as transcriptome analysis, viral genomics, expression analysis, and phylogenetic analysis. The following section provides broad introductory descriptions of these analytical techniques that are used in the four chapters of this thesis; more detailed descriptions of specific material and methods can be found in the original manuscripts.

Study Species

Formica exsecta and other Formica species

The thesis work focuses exclusively on the monogyne (single queen) form of *F. exsecta* collected from five islands close to the Tvärminne Zoological Station in Hanko on the southwest coast of Finland. This population has been surveyed since 1994, for the study of demography, colony kin structure, colony size, productivity, and sex ratio (Sundström et al. 2003; Haag-Liautard 2009; Vitikainen et al. 2011). The wood ant *F. exsecta* (Formicidae; Hymenoptera) is a common Palearctic species and also an important model for the study of social evolution and population genetics (Sundström, Keller & Chapuisat, 2003b; Seppä et al., 2004; Pamilo et al., 2005; Sundström, Seppä & Pamilo, 2005). The colonies produce new sexuals – males and queens – yearly, and in June-July the sexuals fly out of the colonies to disperse and mate on a mating flight, where queens store the sperm from one or several males in their spermatheca; males die after the mating flight. The mounds are predominantly located near the forest edges and in open patches of mixed and deciduous forest. The nests with diameters and depths up to 1–1.5 m, comprising a soil core, overlaid by needles, grass, and small sticks. Nests that reach maturity are relatively long lived (up to 20 years); however, many fail to reach maturity and thereby the average life span is six to seven years (Pamilo, 1991; Haag-Liautard 2009).

In chapter I, altogether 105 individuals from 56 colonies of *F. exsecta* were collected to cover both sexes, female castes, and life stages. In chapter II, 200 adult males were collected from one single-queen colony, which means that a pool of males in altogether are representative of the diploid genome of their mother. The studies in chapters III and IV were carried out using the same datasets as in chapter I. However, for chapter IV, newly generated sequencing data from seven *Formica* ant species

(*F. pressilabris*, *F.fusca*, *F.cinerea*, *F. aquilonia*, *F. truncorum*, *F.pratensis* and *F. exsecta*) were used (Morandin et al., 2016). The *Formica* genus, commonly known as wood ants and mostly live in forest habitats, encompasses 175 described species (Dlussky 1967; Bolton 1995; Goropashnaya et al. 2004). This *Formica* species populations have been studied for *Wolbachia* pathogen screening and transmission dynamics (Viljakainen, Reuter & Pamilo, 2008).

Meta-Transcriptomics Analysis

RNA extractions & sequencing

In order to avoid RNA degradation, the samples relating to Chapter I, III, and IV were stored directly in a -80 ºC freezer, immediately after data collection. For managing the qualitative aspect of the experiment the samples were cleaned from visible exogenous material under a preparation microscope and strictly adhered to the manufacturer's protocol by extracting the total RNA from whole-ground ants individually in TriSure (Bioline, London, UK) as well as removing the contaminated genomic DNA by DNAse I digestion (Fermentas). As the total RNA pools were DNAse-treated, therefore by using poly-A-tail selection they were selected for mRNA to go through the subsequent process of fragmentation. However a double-stranded cDNA synthesis protocol was performed prior to library preparations. For library construction approximately 200 base insert length fragments were selected. Further, in two sets paired-end libraries and sequencing were conducted. Finnish Institute for Molecular Medicine (PE-99) conducted the first set of libraries, whereas the second one by Beijing Genomic Institute BGI (PE-91). For cluster generation, the cBot-2 system and TruSeq PE Cluster Kit v3 (Illumina, San Diego, CA, USA) were used, while TruSeq SBS Kit v3-HS reagent kit and HiSeq2000 instrument (Illumina, San Diego, CA, USA) were used for generating paired data to utilize them in raw read data processing. In the Chapter I the detailed procedure of RNA extraction, library preparation and sample used in specific library were included.

Transcriptome assembly

In Chapter I the detailed workflow of the assembly process is outlined. From the Table 1 in Chapter I it can be summarized as a process of combining multiple library raw data files, extracting from different sexes, life stages, and castes. FastQC (Andrews S., 2010) helped in assessing the quality of raw data and trimmed of the reads to equal length at the 3'-ends became possible with the help of FastX toolkit Version 0.0.13 (Hannon, 2010). After trimming, all high quality reads (average >=20

Q for all base positions on Phred scale) were free from Illumina adapter contamination and primer dimers, with resulting in paired data without any orphan reads. After that, assembling of trimmed reads was done in four different stages of TA such as Initial-TA, Meta-TA, Evidence-TA, and Unigene-TA. However the assessment of transcriptome assemblies was carried out by using multiple methods, such as BUSCO analysis, examining full-length coverage with the closely related organism, and computing contigs assembly statistics. Moreover the functional annotation of transcriptome assemblies was carried out by using the homology search for known sequence database (NR, SwissProt), protein domain identification (PFAM), ontology terms, enzyme information, and information of the protein structure and family.

Microbiome analysis

Identification of microbiota

For identification of microbiota the final meta-transcriptome is used in Chapters III and IV. If the contigs did not match to any ant genomes, then they were searched for homology (BLASTx, version 2.2.26+) against the fungal, virus, and bacterial genome databases at NCBI database (Updated 2015, 2017). Moreover the expression values of the resulting matches were calculated by aligning and counting all the reads from each library, with the help of both FIMM and BGI data. In order to obtain a more accurate net outcome, ribosomal NCBI matches extracted from the transcriptome, namely the 16S, 18S, and 28S ribosomal subunit genes, for analysis (As a cut off- alignment length of 100nt, E-value 0.001, and sequence identity 70%). Additionally, sequence homology searches carried out by using the expressed protein sequences of viruses (DNA and RNA viruses) from the NCBI virus database (Updated 2015,2017) for detecting the whole viral genome sequences.

Meta-Genomics Analysis

DNA extraction & sequencing

DNA extraction was done from testis of 200 adult males of *F. exsecta*, containing sperm cells and organ tissue, for avoiding contamination by gut microbiota. DNA extraction was carried out by using Qiagen Genomic-tip 20/G extraction kit according to the manufacturer's protocol, . Hence we constructed three small insert paired-end libraries (insert sizes of 200 bp, 500 bp, 800 bp) and four mate pair (large insert paired-end) libraries (insert sizes of 2 KB, 5 KB, 10 KB and 20 KB) for Illumina sequencing where each of them contained DNA from 15-50 pooled males. The libraries

were prepared by using the manufacturer's protocol. Moreover sequencing was done at the Beijing Genomics Institute (BGI) by using HiSeq2000, which produced a total of 99.97 GB of raw data.

Genome assembly

By using SOAPdenovo2 version 2.04 (Luo et al., 2012), the *F. exsecta* genome assembly was carried out in three main steps, namely contig assembly, scaffold construction and gap closing. However to identify bacterial contamination, blobology v1.0 (Kumar et al., 2013) software was used to generate taxon-annotated GC-coverage (TAGC) plots of scaffolds in the genome *F.exsecta* assembly. From this analysis it was revealed that 74 scaffolds matched the endosymbiotic bacterium *Wolbachia.* All these 69 scaffolds represented 3.09 MB total, with an N50 value of 104,167 bp, and referred to as "the *Wolbachia* endosymbiont genome of *F. exsecta*" (*w*Fex). However, remaining five contigs were retained in the final assembly for *F. exsecta* as they contained both *Wolbachia* and ant sequences. Following this curation, the final draft genome assembly was 277.7 MB long with an N50 value of 997,654 bp and 36% Guanine-cytosine (GC) content. The quantitative assessment of genome assemblies was carried out by using CEGMA (Parra, Bradnam & Korf, 2007), BUSCA (Simão et al., 2015) datasets, and transcriptome assembly of *F. exsecta*.

Genome Annotation

In order to establish an Official Gene Set (OGS) for the *F. exsecta* genome several publicly available data sets and computational gene prediction tools were combined. The gene sets and gene models from MAKER (Cantarel et al., 2008; Holt & Yandell, 2011) and from other programs like Augustus (Stanke & Morgenstern, 2005) and Glimmer (Salzberg et al., 1998) were then merged with subsequently removing the redundancy.

Moreover, we created final annotation by fetching public database (NR, SwissProt) and collecting information, including gene sequences with retrieving protein-related names, functional domains, and expression in any other organisms along with enzyme commission (EC) numbers, pathway information, Cluster of Orthologous Groups (COG), functional classes, and Gene Ontology terms.

Phylogenetic Analysis

Bacteria Phylogeny

By using the best match cultivated type strains from the Ribosomal Database (http://rdp.cme.msu.edu/) and aligning contigs sequences with those found from the transcriptome, the 16s ribosomal RNA phylogeny was constructed in Chapter III with usage of muscle as implemented in SeaView v. 4.4.1 (Gouy, Guindon & Gascuel, 2010). Further, a phylogenetic tree was constructed by PhyML in accordance with the implementation in SeaView v. 4.4.1 (model choice GTR, 100 bootstraps).

Chapter II describes the analysis process of the *w*Fex phylogeny in MrBayes v3.2.6 x64 (Ronquist & Huelsenbeck, 2003) by using a concatenated sequence of 12 genes, which were subsequently presented as a single copy in *w*Fex genome. However, for this analysis, each gene was considered as a different partition, and the most fitting nucleotide substitution model was selected for each gene, by using the Bayesian information criterion (BIC) in the program jMODELTEST (Posada, 2008). After the creation of the phylogenetic tree it was visualized by using Figtree v1.4.2 (Rambaut, 2012).

Virus phylogeny

Due to the divergence and different genome organizations of the viruses the virus phylogenies were separately constructed in Chapter IV. It was observed that at first the full FeV1 genome sequences and then the partial FeV1-like sequences were aligned to the Dicistroviridae family viruses. In the second step the corresponding data for FeV2 and FeV2-like sequences were aligned to the Iflaviridae family viruses. In the final step the focus was on the third prospective virus from *F. exsecta*, and the other new types of *Formica* species and as such no prior knowledge of its phylogeny could be obtained, except its belonging to the order Mononegavirales. Alignments were carried out at the nucleotide level for FeV1 & FeV2 and protein level for FeV4 by using the software MAFFT (Katoh et al., 2002), while subsequently missing residues were indicated as “N” for all FeV-like short fragments. For each virus, we also constructed separate phylogenomic trees by using the Maximum Likelihood program RAxML (Stamatakis, 2014) under a heuristic approach and the GTR substitution model. The branch supports for the tree topologies were assessed by the bootstrap analysis with 1000 pseudoreplicates of the sequences. The figures were created by using FIGTREE version 1.2 (Rambaut, 2012).

Main results and their interpretation

Insights from transcriptome profiling of *F. exsecta*

In order to maximize the representation of the expressed genes, we explored the overall transcriptome profiling of *F. exsecta* in Chapter I, and then the RNA sequencing libraries were constructed from pooled samples of several life stages of both sexes as well as female castes of queens and workers. After reviewing the quality filter analysis, we obtained 322 million high quality pair-end reads in aggregate, which were then used to construct initial assemblies from three transcriptome assemblers (Trinity, SOAPdenovoTrans and Velvet-Oases). However, depending on the software, the transcriptome assemblers reported the total assembly lengths of initial assemblies as up to 85–245 million bases (MB) with the maximum number of contigs generally varies between ca. 200 k and 300 k.

The assessment of the initial transcriptome assemblies from the multiple transcriptome assemblers revealed that Trinity assembler performed the best for our data, and produced 7,145 full length transcripts with more than 90% coverage of CflorPP proteins. I achieved the highest-quality transcriptome (Final-TA) by combining the output from the Trinity de novo assembler with different k-mers (range: 21–31) and presented the simultaneous assessment of a variety of metrics.

Final assembly was constructed with the help of the combined data (pooling males, queens and workers), covering nearly 7,435 *C. floridanus* genes (with 90% of their protein length), whereas separately, an assembly of males, queens and workers, each covering fewer genes with 90% coverage cut-off (4,215, 6,134, and 5,913 genes, respectively Figure 2A).

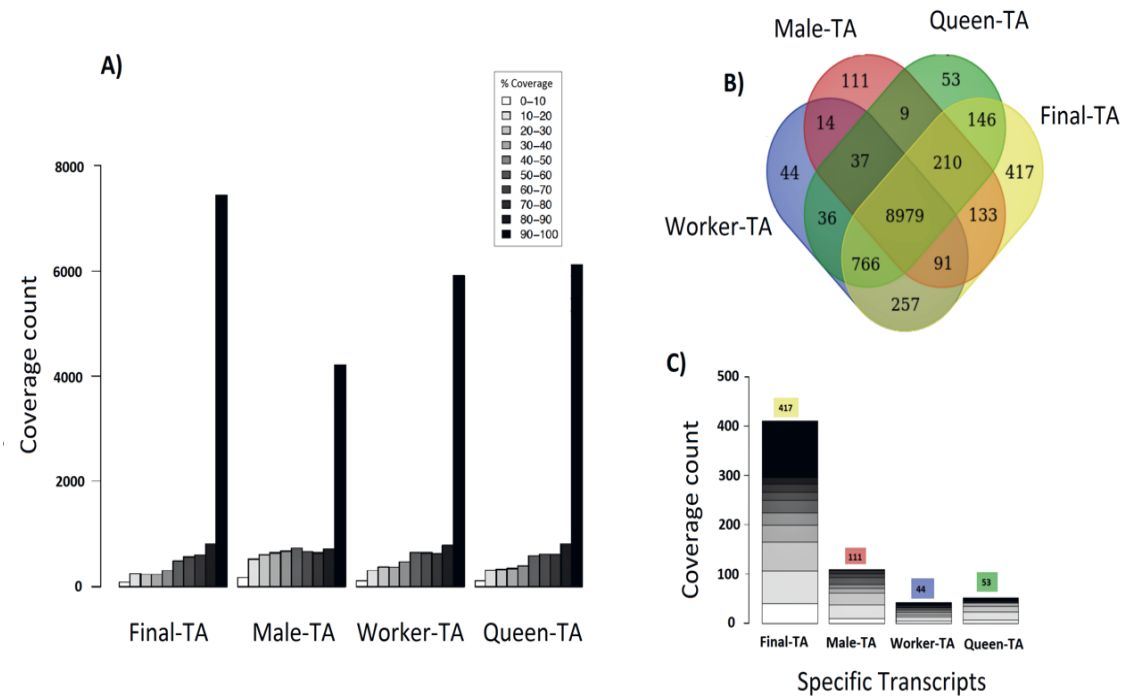

Figure 2: Coverage of contigs assembly between castes.

(A) Alignment of caste-specific assemblies to *C. floridanus* predicted proteins compared to the Final-TA, including all castes. (B) Overlap of transcripts aligning to *C. floridanus* proteins, by castes and the Final-TA. (C) Alignment coverage of contigs to *C. floridanus* assembled by caste or in the Final-TA.

In total, in Chapter I, the 42,476 contigs in the final transcriptome were annotated with 17,496 unique genes across all the 31 ant species available in the NCBI database, and 8,906 proteins top matches came from *C. floridanus.* The total size of the annotated contigs was 87 MB base pairs, whereas the unannotated ones only comprised 37 MB base pairs, despite being more numerous.

This suggested that many of the unannotated sequences were fragmented. Additionally, the comparative analysis of the final transcriptome of *F. exsecta* and the *C. floridanus* genome revealed that 3,087 genes contained several isoforms (range 2–84, average = 4.1, s.d. = 4.2). Comparatively, in *Drosophila melanogaster,* approximately 60% of the multi-exon genes were estimated to contain several isoforms (Stolc et al., 2004).

A) Caste-specific expressed transcripts

Separate assemblies of the castes showed that 8,979 *C. floridanus* genes (3,725 of which had 90–100% coverage, Figure 2A), were shared across the castes (Figure 2B), whereas 536 genes were unique to a caste or sex (133 in males, 257 in workers, and 146 in queens), and 766 were specifics to the two female castes combined (Figure 2C). A comparison of assemblies from different castes (male, queen, worker) revealed that the combined assembly had 417 castes-specific gene transcripts (Figure 2C). These results suggest that analyses relying on single life stages or morphs are likely to miss some genes due to the low sequencing depth or expression at specific stages.

B) Expression profile of gene families

The evolution of many gene families is thought to be affected by the evolution of eusociality (Simola et al., 2013). Gene families relating to chemical communication (Ban et al., 2003; Kulmuni & Havukainen, 2013), immune system (Evans et al., 2006; Cremer, Armitage & Schmid-Hempel, 2007; Viljakainen et al., 2009; Roux et al., 2014) and caste differentiation (Wahli et al., 1981; Schwander et al., 2010; Feldmeyer, Elsner & Foitzik, 2014) have been predicted to be diversified in social insects, as compared to solitary insects. The final transcriptome in Chapter I reported five copies of Vg and one Vg receptor, as well as their isoforms Vitellogenin-like-B (four isoforms), Vitellogenin-like-D (two isoforms), Vitellogenin receptor (four), and only one isoform with each of Vitellogenin-like-A, Vitellogenin-like-C, and the conventional Vitellogenin genes. None of these were specific to any sex or caste. Earlier work on *F. exsecta*, where the same reads of queens and workers were applied, found altogether four vitellogenin gene orthologues, (Vitellogenin, Vitellogenin-like-A, Vitellogenin-like-B, and Vitellogenin-like-C), each with specific expression patterns and great structural variation (Morandin et al., 2014).

Final-TA assembly in Chapter I observed 12 chemosensory protein (CSP) genes, and three odorant binding protein (OBP) genes, in which the number of isoforms varied between 1 to 10 (average = 2.2, s.d. = 2.0). The ranges of this observation have similarities with other species of ants, in which the number of functional CSP genes ranges from 11 to 21 depending on the species (Vieira & Rozas, 2011; Kulmuni & Havukainen, 2013). In addition, our transcriptome contained only some of the antimicrobial-peptide coding genes, described previously in ant genomes (77 unique genes, up to 10 isoforms, average = 1.4, s.d. = 1.4). So, any conclusion about the absence of genes from the genome and gene family sizes in general should be treated with caution when based on expression data only.

C) Microbial community diversity

By analyzing data of non-ant origin from transcriptome assemblies in Chapter III, the identity sequencing of a wide range of microbes and some species of mite were retrieved from within and on the ant *F. exsecta.* However the Chapter III findings included members of several endogenous bacterial phyla, including *Wolbachia*, two obligate endogenous, possibly entomopathogenic fungi, and the complete genomes of three novel RNA viruses belonging to the classes of Iflaviridae, Dicistroviridae, and Mononegaviridae. Since transcriptome data are generated from the whole RNA, thereby the sequence matches most likely represent some of the most active microbes associated with *F. exsecta*. Sequences matching *Wolbachia* showed high levels of identity with two strains that commonly infect Eurasian and Finnish populations of several closely related *Formica* wood ants, including *F. exsecta*. In the related *Formica rufa* one strain of *Wolbachia* infect near 100% of colonies, with the second strain occurring at lower frequencies. The transcriptome data suggest a similar relationship between the strains in *F.exsecta*. In addition to sequences yielding identity to *Wolbachia*, there exist sequences with homology to other intracellular microbes. These were to the fungi Microsporidia, and to bacteria belonging to the Enterobacteriacae (Arsenophonus), the Entomoplasmatales, and the Aceterobacteraceae (Saccharibacter). Several 16S rRNA sequences had their nearest GenBank match to Aceterobacteraceae bacteria. Aceterobacteraceae have been found in honey bee guts, and one such bacteria, Saccharibacter, has also been isolated from pollen, and may hence represent bacteria found through foraging.

Insight from *Formica exsecta* genome

In order to test whether horizontally transferred genetic elements exist in the genome of the ant *F. exsecta* and to describe the genomic organization of any such elements, this thesis (Chapter II) aims to produce the first genome assembly of the wood ant *F. exsecta.* Not only due to the key taxonomic position, but also on the basis of the available ecological and behavioral data, the genus *Formica* is listed by the Global Ant Genome Alliance (GAGA) as one of the high-priority ant taxons to be sequenced. The Illumina sequencing libraries from DNA, extracted from the testes of males of a *F. exsecta* colony yielded >99 gigabases of Illumina sequence data.

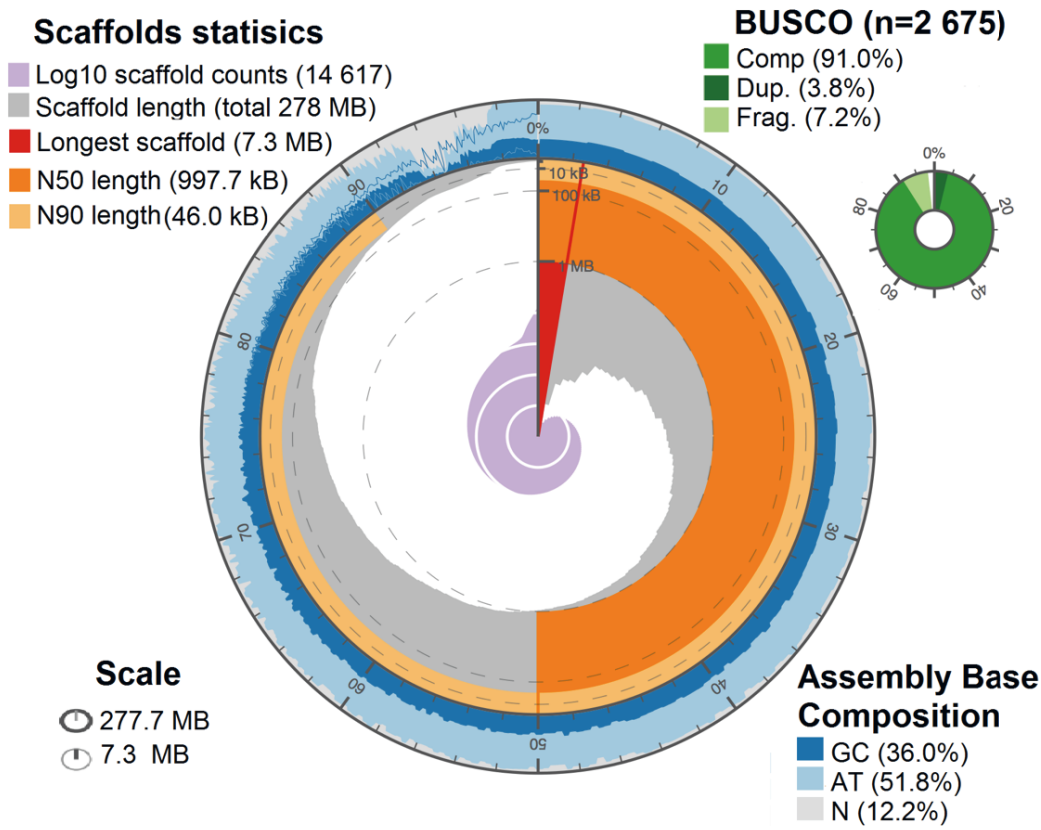

Figure. 3 De novo genome assembly of *F. exsecta* genome, summarized by the following metrics: a) Overall assembly length, b) Number of scaffolds/contigs, c) Length of the longest scaffold/contig, d) Scaffold/contig N50 and N90, e) Percentage GCs and percentage Ns, f) BUSCO completeness, g) Scaffold/contig length/count distribution

The final genome resulting from the assembly of these data was 277.7 megabases (MB) long, encompassing 14,617 scaffolds (Figure 3) with an N50 scaffold length of 997.7 KB. The number of scaffolds is higher than the number of chromosomes (n=26) reported for *F. exsecta*. Similarly, the *F. exsecta* genome assembly in Chapter II is somewhat shorter than the genome size estimates obtained by flow cytometry for species in the subfamily Formicinae (range: 296-385 MB). I identify 13,767 protein coding genes for which I provide gene ontology and protein domain annotations. The raw data, gapped scaffolds, and annotations underpinning this assembly are deposited in public databases under BioProject PRJNA393850 (accession NPMM00000000). The *F. exsecta* genome assembly is

comparable in terms of quality as well as completeness with other sequenced ant genomes. All the 248 CEGMA eukaryotic core genes were found and out of them 241 genes were complete in length. Similarly, 98.5% of the 1634 BUSCO Insecta genes were complete in the genome. More than 98.75 % of the 10,999 assembled ESTs were mapped unambiguously to the genome. Altogether, these analyses show that the genome assembly has high completeness.

A) Horizontal gene transfers, and functional novelty

Intracellular symbionts can contribute new genes or fragments of genes to the host genome via horizontal gene transfer (HGT). In this thesis study (Chapter II), I found evidence for ancestral horizontal transfer of cytoplasmic *Wolbachia* to the host *F. exsecta* in five scaffolds (scaffold83, scaffold233, scaffold574, scaffold707, scaffold741) (chromosomal *Wolbachia*). The four largest transfers are 13 to 47 kb long, and include 83 putative functional protein coding genes, whereas the fifth and smallest insertion (475 bp) lacks protein coding genes, other than a degenerate *Wolbachia* transposase. This transposase is present in 7 out of 29 published *Wolbachia* genomes. The chromosomal *Wolbachia* showed high similarity to the cytoplasmic *Wolbachia* (88.2 – 99.2 %). Most of these transferred fragments contained transposable elements, as well as some other functional genes from the *Wolbachia* genome. The presumptive HGT events from *Wolbachia* to *F. exsecta* are located in or near regions with transposases. Whether the presence of such transposases close to HGT sites facilitates insertions is unknown. Interestingly, the putative functional protein-coding genes of *Wolbachia* inserted in the *F. exsecta* genome are similar to the genes reported in similar HGTs events in other insect genomes (eg: ABC transporter, Ankyrin repeat containing protein. This could indicate that some HGT events are either more likely to occur or to be retained for reasons that could be neutral or adaptive to the host or to the endosymbiont.

B) Valuable resources for future comparative genomics

As *F. exsecta* genome (chapter II) is first genome released from genus *Formica* and is expected to be a valuable resource in understanding the molecular basis of the evolution of social organization in ants. Recent genomic comparisons between *Formica selysi* and *Solenopsis invicta* have shown convergent evolution of a social chromosome, that underpins social organisation in these ants. Additional comparison of these genomic regions with *F. exsecta* could provide valuable insights on the evolution of genomic architectures underlying social organization. In this thesis (Chapter II), comparative analysis of the *F. exsecta* genes with the closely related species *C. floridanus* and *L.*

niger, and the more distantly related *S. invicta* and *C. biroi* revealed, that 4,685 orthologous clusters out of 7,727 are shared between all five. In addition, I found 102 gene clusters that were exclusive to three Formicinae genomes (*F. exsecta*, *Camponotus floridanus* and *Lasius niger)* Such genes are important candidates that could be involved in the evolution of this subfamily. Also, selection analysis (dN/dS ratio estimations) on 3,157 single-copy genes shared between the five core ant species (without paralogous genes), revealed that 500 genes have signatures of positive selection in the lineage leading to *F. exsecta*. These include genes involved in fatty acid metabolism, lipid catabolism, and chitin metabolism. Interestingly, previous studies on ants, bees, and flies also provide evidence for positive selection on genes in similar functional categories as in our study. The modalities of the putative faster evolution of these genes will become clearer as genome sequence becomes available from other Formicinae.

Insights from *w*Fex genome

The assembly of the "*Wolbachia* endosymbiont genome of *F. exsecta*" (henceforth *w*Fex), was 3.09 Mb long, encompassing 69 scaffolds with a N50 scaffold length of 104,167 nt, and a GC content of 35.13% (Table 1; GenBank, Bioproject: PRJNA436771). This assembly of *w*Fex shows extensive nucleotide similarity with the *Wolbachia* endosymbiont of *Dactylopius coccus*, strain *w*DacA (GenBank ID: NZ_LSYX00000000). However, the *w*Fex genome is considerably larger (3.09 Mb) than the *Wolbachia* genomes reported previously (range: 0.95 to 1.66 Mb), and includes a greater number of open reading frames (1,796 ORFs) than other published *Wolbachia* genomes [range: 644 to 1,275 genes]. In the phylogenetic analysis, *w*Fex clustered with the *Wolbachia* strains within super group A. This is consistent with earlier studies on *Wolbachia* in ants, which also found supergroup A in the majority of the infected ants. The closest relative of *w*Fex was the strain *w*DacA which infects the scale insect, *Dactylopius coccus* (Figure 4). Our phylogeny is also consistent with the recent published phylogeny of *Wolbachia.* To explore differences in gene content between CI-inducing, and non-CI-inducing strains of *Wolbachia*, homologous genes in six CI-inducing, and three non-CI-inducing strains were aligned, and compared. The CI-inducing strains shared 84 genes, not found in the non-CI-inducing strains. We found 80 (95.23%) of these 84 genes in *w*Fex as well as the genes *cifA* and *cifB*, which are involved in the CI mechanism. Both copies of genes appear to be functional as their lengths are 100% in comparison to similar gene sequences available in NCBI database. Together this supports the assumption that *w*Fex is a CI-inducing *Wolbachia* strain, but we warrant that genomic information is unable to conclusively demonstrate this.

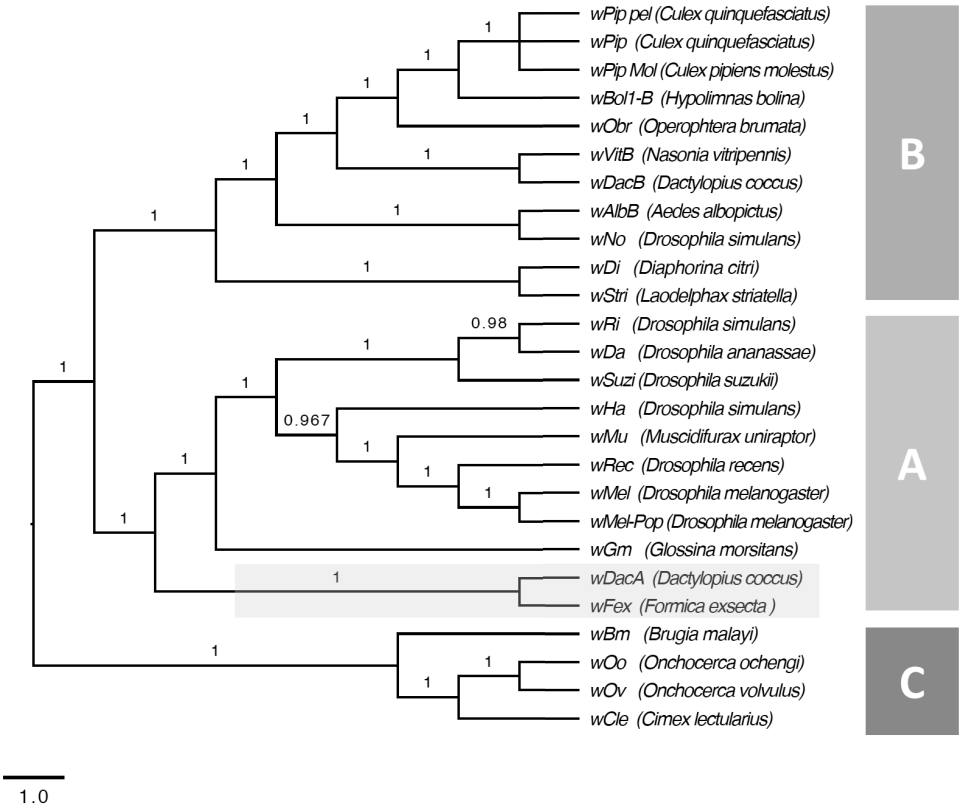

Figure 4. Phylogeny of the *Wolbachia* supergroups a, b, and c strains with the newly assembled *w*Fex genome. The phylogenetic reconstructions are based on individual analyses of 12 single copy core genes of 25 *Wolbachia* strains. The support values on the branch labels indicate Bayesian posterior probabilities. The letters a-c indicates the separate supergroups

Nature of FeV1, FeV2, and FeV4 viruses

The meta-transcriptomic data of *F. exsecta* reveals three probable full virus genomes (Chapter IV) and examine the genome organization and molecular characterization, along with ORF predictions, and functional annotation of genes. The *Formica exsecta virus-4* (FeV4; GenBank ID: MF287670) is a newly discovered negative-sense single-stranded RNA virus representing the first identified member of order Mononegavirales in ants, whereas the *Formica exsecta virus-1* (FeV1; GenBank ID: KF500001), and the *Formica exsecta virus-2* (FeV2; GenBank ID: KF500002) are positive

single-stranded RNA viruses (Figure 5),. The average genome coverage was 6900X for virus FeV1, 3020X for FeV2 and 2651X for FeV4 (Figure 5), in the *F. exsecta* transcriptome (Chapter I).

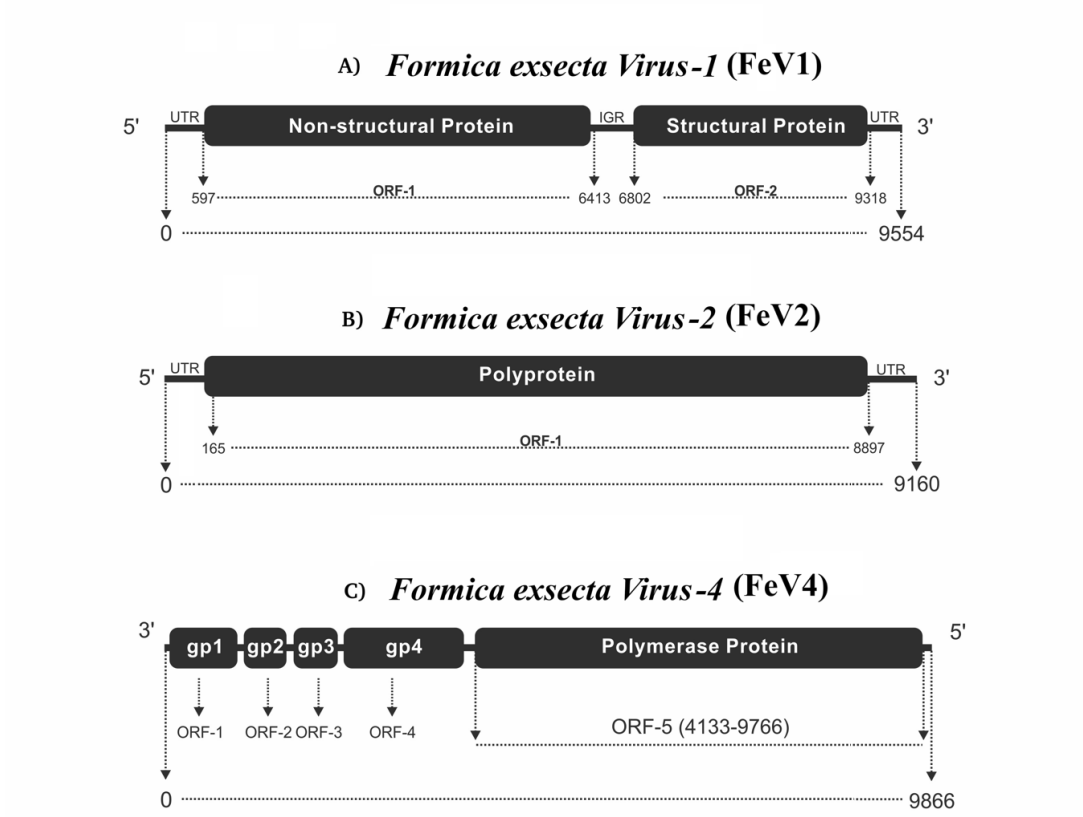

Figure 5: Comparative genome architecture of the three *F. exsecta* viruses FeV1 (A), FeV2 (B), and FeV4 (C). The figure shows the orientation of the genome, the horizontal black bars indicate the number of cistrons, and the text in them the type of protein. The relative positions of ORFs are indicated by rectangles.

They exhibit differences in genome organization; FeV1 is dicistronic, FeV2 is monocistronic, whereas FeV4 is pentacistronic. FeV1 and FeV2 show similarity in sequence and genome organization to many positive-strand RNA viruses that infect ants and honey bees (Chen & Siede, 2007; Valles, 2012; Sébastien et al., 2015). These viruses were found in all castes, and age classes of *F. exsecta*. All three viruses were also detected in the new field-collected mature workers. Some colonies were infected by multiple viruses, and the viruses were observed to infect all castes, and multiple life stages of workers and queens. The three viruses are phylogenetically distinct, and

according to our phylogenetic analysis, group into three separate virus families (Dicistroviridae, Iflaviridae, and Nyamiviridae, respectively). Finally, the results also show that some other ant species (*F. fusca, F. pressilabris, F. pratensis, F. aquilonia, F. truncorum and F. cinerea*) of the genus *Formica* contain sequence fragments with close phylogenetic affinity to the three FeV viruses, but these results will need further revalidation.

Conclusions

This thesis work encompasses four studies aimed at increasing our knowledge of microbial community associated with ants, and understanding the link between microbe diversity, community structure, and host-microbe interactions. The study presents the first draft genome of the ant *F. exsecta*, and its *Wolbachia* endosymbiont. Genome organization results such as HGT, gene duplications, isoforms suggest that draft genomes can be used as a reference for ecological genomics or genetic studies of *Formica* ants related species and understanding host-parasite interactions between them. In addition, the first report of a *Wolbachia* draft genome could provide a valuable resource for resolving *Wolbachia* transmission dynamics in ants.

The metagenomics and metatranscriptomics techniques, also enabled us to gain a better insight in host microbial communities. High-throughput sequencing revealed that probable microbial communities such as intracellular bacterial, soil bacterial, and fungal communities and viruses that have coevolved and diversified with *F. exsecta* ant. In general, understanding microbial community will unquestionably offer novel insights towards understanding crucial unresolved aspects of hosts and microbes in symbiosis. Therefore, the role of big data and multi-omic approaches in elucidating dynamics of microbial ecosystems cannot be underemphasized, as they have the capability to redefine our understanding of how ecosystems respond and feed back to produce system change.

Acknowledgements

For allowing me to be here today, I am thankful to so many people who have encouraged, supported and helped me through this long voyage.

I would like to start by thanking my supervisor prof. Lotta Sundström. Lotta, I am so grateful to you for giving me this opportunity. The value of this opportunity cannot be expressed in words alone. You have helped me to grow as a scientist and a person, and have given me a whole new appreciation for doing science. You have been a great mentor and I have always felt you are there for me when I needed it. Thanks for all the advice and the resource to carry out such bold experiments. I would also like to thank my second supervisor Helena for guiding me, and for your support during times of my PhD. I am thankful to you for your valuable suggestions and for cheerful comments. I am grateful Juan for leaving me the freedom, for fruitful collaboration, and for sharing an office in Jyväskylä.

I wish to thank Pekka Pamilo and Ari Loytynoja, for being the members of my thesis committee and accompanying me with cheerful comments throughout the years. Also, thanks to the referees of this thesis, for taking some of their precious time to review my thesis work.

The most time during my PhD were spend in Viiki with Antzz group. I really want to thank everybody in the Antzz group; all of you have somehow contributed to this thesis. I want to thank Jonna for always taking the time to comment on my presentations, manuscript, for discussing projects. Thanks to Stafva for sharing the office, for giving me motivations, sharing the gossip and helping me to deal with the administrative work at various stages. Sanja, Janna, Dimi, Jack, Unni, Jenni, Anton, Martina, Dalial, Nick, Claire, Siiri, Heli, Rose, Perttu, Heikki, Kalle and many others thanks to all of you for discussing science, for the fun times, many parties, conference trips and Monday meetings. Minttu, Heini, Leena and Leila for the great help in the MES lab.

During my PhD work, I have had the opportunity to travel lot and meet remarkable people. I want to thank Yannick for his fruitful collaboration, interesting discussions, new ideas and good times in Queen Marry University, London. Furthermore, I want to thanks CoE people at University of Jyväskylä for discussing science. Special thanks to Prof. Johanna Mappes for giving me the opportunity to work in CoE, and for all help during Jyväskylä visits. Thanks to all the others with whom I shared a drink or a conversation. I extend my gratitude to my teachers from University of

Pune, Prof. Mohan Kale and Prof. Urmila Kulkarni, for motivating and inspiring me to pursue PhD studies.

The Faculty of Biological and Environmental Sciences have been an excellent place to work. I would also like to thank Veijo Kaitala and Perttu Seppä for help with the administrative formalities. I am also thankful to LUOVA coordinators Anni Tonteri, Anita and Karen, for helping me and advising me in my studies, travel grants requirements. Special thanks to LUOVA for many travel grants given to me for travels abroad for workshops and conference. Thanks to all LUOVA PhD students and staff for Wednesday seminars, coffee breaks, and discussing science. My non-ANTZZ friends from various groups- Thank you Sergey, Cui Wang, Baocheng, Scott for your friendship, helpful discussions and some relaxing times. I would like to thank my Prof. Karl Lemström and Sanna Salmi for giving me new role in Transplantation laboratory and for enormous support. Thanks to all new colleagues of Transplantation laboratory for welcoming me in the group and a great time.

My time in Finland would definitely not be the same without Abhilash and Himanshu. Thank you for all the time we have spent together, for your friendship and for always being there for me. Many thanks to my Indian friends in Helsinki: Kiran, Swapnil, Ashwini Kumar, Bhupendra, Sarang, Bharat, Sushil Tripathi, Mahesh, Om, Vishal, Amit, Dipti, Sreesha and many others for fun-n-food filled get-togethers. Thank you for your companionship and support, especially for the cricket matches during short summers of Helsinki. I am grateful to my IT friends Vitthal Wable, Jyoti Wable, Sushil Pawar, Ashwin Rao, Aniruddha, Swapnil and Dhruv Anand for helping in keeping my mind off work and studies.

Finally, and most importantly, I want to say thank you to everyone in my family for their support and love to me. I thank my father and mother for extending their support and encouragement during my stay away from home. Thank you Ramchandra Dhaygude, Mahendra, Vinod, Hanumant, Archana and all other Dhaygude family members for your enormous support to me. I am most grateful to my wife Prajkta, for being understanding, for your great help, and for loving me. I am the luckiest father in the world for having Kripa, my little scientist, at home when I come from work. Thank you Kripa for bringing lots of happiness to my life. I love you so much.

माझे शैक्षणिक यश हे वडिलांची इच्छा आणि आईचे स्वप्न होते. आई-वडिलांचे कष्ट हीच प्रेरणा उराशी बाळगत ध्येयासाठी झटत होतो. मला खूप आनंद होत आहे, मी आज आई-वडिलांचे स्वप्नंपूर्ण केले. I dedicate PhD thesis to my parents.

References

Ban L., Scaloni A., Brandazza A., Angeli S., Zhang L., Yan Y., Pelosi P. 2003. Chemosensory proteins of *Locusta migratoria*. *Insect Molecular Biology* 12:125–134. DOI: 10.1046/j.1365-2583.2003.00394.x.

Boomsma JJ., Brady SG., Dunn RR., Gadau J., Heinze J., Keller L., Moreau CS., Sanders NJ., Schrader L., Schultz TR., Sundström L., Ward PS., Wcislo WT., Zhang G. 2017. The Global Ant Genomics Alliance (GAGA). *Myrmecological News* 25:61–66.

Brown W., Liautard C., Keller L. 2003. Sex-ratio dependent execution of queens in polygynous colonies of the ant *Formica exsecta*. *Oecologia* 134:12–17. DOI: 10.1007/s00442-002-1072-8.

Cantarel BL., Korf I., Robb SMC., Parra G., Ross E., Moore B., Holt C., Sánchez Alvarado A., Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 18:188–96. DOI: 10.1101/gr.6743907.

Chen YP., Siede R. 2007. Honey Bee Viruses. *Advances in Virus Research* 70:33–80. DOI: 10.1016/S0065-3527(07)70002-7.

Conte Y Le., Hefetz A. 2008. Primer Pheromones in Social Hymenoptera. *Annual Review of Entomology* 53:523–542. DOI: 10.1146/annurev.ento.52.110405.091434.

Correa CC., Ballard JWO. 2016. *Wolbachia* Associations with Insects: Winning or Losing Against a Master Manipulator. *Frontiers in Ecology and Evolution* 3:153. DOI: 10.3389/fevo.2015.00153.

Cremer S., Armitage SAO., Schmid-Hempel P. 2007. Social Immunity. *Current Biology* 17:R693–R702. DOI: 10.1016/j.cub.2007.06.008.

Deslippe R. 2010. Social Parasitism in Ants. *Nature Education Knowledge* 1:27. DOI: 10.1111/j.1365-294X.2005.02499.x.

Douglas AE. 2011. Lessons from studying insect symbioses. *Cell host & microbe* 10:359–67. DOI: 10.1016/j.chom.2011.09.001.

Engel P., Moran NA. 2013. The gut microbiota of insects-diversity in structure and function. DOI: 10.1111/1574-6976.12025.

Evans JD., Aronstein K., Chen YP., Hetru C., Imler J-L., Jiang H., Kanost M., Thompson GJ., Zou Z., Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees Apis mellifera. *Insect molecular biology* 15:645–56. DOI: 10.1111/j.1365-2583.2006.00682.x.

Fanning S., Proos S., Jordan K., Srikumar S. 2017. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology* 8:1–16. DOI: 10.3389/fmicb.2017.01829.

Feldmeyer B., Elsner D., Foitzik S. 2014. Gene expression patterns associated with caste and reproductive status in ants: worker-specific genes are more derived than queen-specific ones. *Molecular Ecology* 23:151–161. DOI: 10.1111/mec.12490.

Ge X., Li Y., Yang X., Zhang H., Zhou P., Zhang Y., Shi Z. 2012. Metagenomic Analysis of Viruses from Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China. *Journal of Virology* 86:4620–4630. DOI: 10.1128/JVI.06671-11.

Gibbons SM., Gilbert JA. 2015. Microbial diversity--exploration of natural ecosystems and microbiomes. *Current opinion in genetics & development* 35:66–72. DOI: 10.1016/j.gde.2015.10.003.

Gouy M., Guindon S., Gascuel O. 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* 27:221–224. DOI: 10.1093/molbev/msp259.

Haapaniemi K., Pamilo P. 2012. Reproductive conflicts in polyandrous and polygynous ant *Formica sanguinea*. *Molecular Ecology* 21:421–430. DOI: 10.1111/j.1365-294X.2011.05386.x.

Hannon. 2010. FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html)).

Hölldobler B., Wilson EO. 1990. *The ants*. Cambridge Mass.: Belknap Press of Harvard University

Press.
Holt C., Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491. DOI: 10.1186/1471-2105-12-491.
Hughes GL., Rasgon JL. 2012. *Wolbachia infections in arthropod hosts*. Elsevier Inc. DOI: 10.1016/B978-0-12-384984-7.00009-9.
Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–66.
Kim M-S., Whon TW., Bae J-W. 2013. Comparative Viral Metagenomics of Environmental Samples from Korea. *Genomics & Informatics* 11:121. DOI: 10.5808/GI.2013.11.3.121.
Krishnan M., Bharathiraja C., Pandiarajan J., Prasanna VA., Rajendhran J., Gunasekaran P. 2014. Insect gut microbiome – An unexploited reserve for biotechnological application. *Asian Pacific Journal of Tropical Biomedicine* 4:S16–S21. DOI: 10.12980/APJTB.4.2014C95.
Kulmuni J., Havukainen H. 2013. Insights into the Evolution of the CSP Gene Family through the Integration of Evolutionary Analysis and Comparative Protein Modeling. *PLoS ONE* 8:e63688. DOI: 10.1371/journal.pone.0063688.
Kumar S., Jones M., Koutsovoulos G., Clarke M., Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in genetics* 4:237. DOI: 10.3389/fgene.2013.00237.
Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung DW., Yiu S-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T-W., Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18. DOI: 10.1186/2047-217X-1-18.
Meunier J. 2015. Social immunity and the evolution of group living in insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370:20140102–20140102. DOI: 10.1098/rstb.2014.0102.
Morandin C., Havukainen H., Kulmuni J., Dhaygude K., Trontti K., Helanterä H. 2014. Not only for egg yolk--functional and evolutionary insights from expression, selection, and structural analyses of *Formica* ant vitellogenins. 31:2181–93. DOI: 10.1093/molbev/msu171.
Morandin C., Tin MMY., Abril S., Gómez C., Pontieri L., Schiøtt M., Sundström L., Tsuji K., Pedersen JS., Helanterä H., Mikheyev AS. 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome biology* 17:43. DOI: 10.1186/s13059-016-0902-7.
Pamilo P. 1991. Life span of queens in the ant *Formica exsecta*. *Insectes Sociaux* 38:111–119. DOI: 10.1007/BF01240961.
Pamilo P., Zhu D., Fortelius W., Rosengren R., Seppä P., Sundström L. 2005. Genetic patchwork of network-building wood ant populations. *undefined*.
Parra G., Bradnam K., Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)* 23:1061–7. DOI: 10.1093/bioinformatics/btm071.
Posada D. 2008. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution* 25:1253–1256. DOI: 10.1093/molbev/msn083.
Prasad RK., Chatterjee S., Sharma S., Mazumder PB., Vairale MG., Raju PS. 2018. Insect Gut Bacteria and Their Potential Application in Degradation of Lignocellulosic Biomass: A Review. In: Springer, Singapore, 277–299. DOI: 10.1007/978-981-10-7485-1_14.
Rambaut A. 2012. Figtree.
Ronquist F., Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574. DOI: 10.1093/bioinformatics/btg180.
Roux J., Privman E., Moretti S., Daub JT., Robinson-Rechavi M., Keller L. 2014. Patterns of Positive

Selection in Seven Ant Genomes. *Molecular Biology and Evolution* 31:1661–1685. DOI: 10.1093/molbev/msu141.

Salzberg SL., Delcher AL., Kasif S., White O. 1998. Microbial gene identification using interpolated Markov models. *Nucleic acids research* 26:544–8.

Schmid-Hempel P. 1995. *Parasites and social insects*. Springer Verlag.

Schmidt TM., DeLong EF., Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology* 173:4371–4378. DOI: 10.1128/jb.173.14.4371-4378.1991.

Schwander T., Lo N., Beekman M., Oldroyd BP., Keller L. 2010. Nature versus nurture in social insect caste differentiation. *Trends in Ecology & Evolution* 25:275–282. DOI: 10.1016/j.tree.2009.12.001.

Sébastien A., Lester PJ., Hall RJ., Wang J., Moore NE., Gruber MAM., Sébastien A., Hall RJ., Wang J., Moore NE., Gruber MAM. 2015. Invasive ants carry novel viruses in their new range and form reservoirs for a honeybee pathogen. *Biology Letters* 11:20150610. DOI: 10.1098/rsbl.2015.0610.

Seppä P., Gyllenstrand N., Corander J., Pamilo P. 2004. Coexistence of the social types: genetic population structure in the ant *Formica exsecta*. *Evolution; international journal of organic evolution* 58:2462–71.

Simão FA., Waterhouse RM., Ioannidis P., Kriventseva E V., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.

Sogin ML., Morrison HG., Huber JA., Mark Welch D., Huse SM., Neal PR., Arrieta JM., Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* 103:12115–20. DOI: 10.1073/pnas.0605127103.

Staley C., Sadowsky MJ. 2016. Application of metagenomics to assess microbial communities in water and other environmental matrices. *Journal of the Marine Biological Association of the United Kingdom* 96:121–129. DOI: 10.1017/S0025315415001496.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30:1312–3. DOI: 10.1093/bioinformatics/btu033.

Stanke M., Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* 33:W465-7. DOI: 10.1093/nar/gki458.

Sundström L., Chapuisat M., Keller L. 1996. Conditional Manipulation of Sex Ratios by Ant Workers: A Test of Kin Selection Theory. *Science* 274:993–995. DOI: 10.1126/science.274.5289.993.

Sundström L., Keller L., Chapuisat M. 2003a. Inbreeding and sex-biased gene flow in the ant *Formica exsecta*. *Evolution; international journal of organic evolution* 57:1552–61.

Sundström L., Keller L., Chapuisat M. 2003b. Inbreeding and sex-biased gene flow in the ant *Formica exsecta*. *Evolution; international journal of organic evolution* 57:1552–61. DOI: 10.1111/j.0014-3820.2003.tb00363.x.

Sundström L., Seppä P., Pamilo P. 2005. *Genetic population structure and dispersal patterns in Formica ants-a review*.

Tringe SG., Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology* 11:442–446. DOI: 10.1016/j.mib.2008.09.011.

Valles SM. 2012. Positive-Strand RNA Viruses Infecting the Red Imported Fire Ant, *Solenopsis invicta*. *Psyche: A Journal of Entomology* 2012:1–14. DOI: 10.1155/2012/821591.

Vieira FG., Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome biology and evolution* 3:476–90. DOI: 10.1093/gbe/evr033.

Viljakainen L., Evans JD., Hasselmann M., Rueppell O., Tingek S., Pamilo P. 2009. Rapid Evolution of Immune Proteins in Social Insects. *Molecular Biology and Evolution* 26:1791–1801. DOI: 10.1093/molbev/msp086.

Viljakainen L., Reuter M., Pamilo P. 2008. *Wolbachia* transmission dynamics in *Formica* wood ants. *BMC evolutionary biology* 8:55. DOI: 10.1186/1471-2148-8-55.

Vitikainen E., Haag-Liautard C., Sundström L. 2011. Inbreeding and reproductive investment in the ant *Formica exsecta*. 65. DOI: 10.1111/j.1558-5646.2011.01273.x.

Wahli W., Dawid IB., Ryffel GU., Weber R. 1981. Vitellogenesis and the vitellogenin gene family. *Science (New York, N.Y.)* 212:298–304.

Weiss B., Aksoy S. 2011. Microbiome influences on insect host vector competence. *Trends in Parasitology* 27:514–522. DOI: 10.1016/J.PT.2011.05.001.

Wilson EO. 1971. *Insect Societies. The Belknap press of Harvard University Press, Cambridge, Massachusetts.*